Routing and Performance Evaluation of Dual Priority Delta Networks under Hotspot Environment

D. C. Vasiliadis, G. E. Rizos, C. Vassilakis
Department of Computer Science and Technology
University of Peloponnese
Tripolis, Greece
dvas@uop.gr, georizos@uop.gr, costas@uop.gr

Abstract—Large swings in the demand for content are commonplace within the Internet. Although Multistage Interconnection Networks (MINs) are fairly flexible in handling varieties of traffic loads, their performance considerably degrades by hotspot traffic, especially at increasing size networks. As alleviation to the tree saturation problem, the prioritizing of packets is proposed leading to a scheme that natively supports multi priority traffic. In this paper the performance evaluation of double-buffered Delta Networks under single hotspot setups, with different offered loads, and 2-class routing traffic is presented and analyzed using simulation experiments. Performance comparison of dual vs. single priority scheme is outlined under hotspot environment, by calculating a universal performance factor, which effectively includes the importance aspect of each of the two most important performance metrics, namely packet throughput and delay. The findings of this paper can be used by MIN designers to optimally configure their networks.

Keywords—Multistage interconnection networks; performance evaluation; hotspot traffic; simulation

I. INTRODUCTION

Multistage Interconnection Networks (MINs) with crossbar Switching Elements (SEs) are frequently proposed as an interconnection infrastructure in parallel multiprocessor systems and network systems alike. In the domain of parallel systems, MINs undertake processor-to-memory communication, whereas in network systems they are typically used in communication devices such as gigabit Ethernet switches, terabit routers, and ATM switches. The significant advantages of MINs are their good performance, their low cost/performance ratio and their ability to route multiple communication tasks concurrently. MINs with the Banyan [1] property, such as Omega Networks [2], Delta Networks [3], and Generalized Cube Networks [4] are generally preferred over non-Banyan MINs, since the latter are -in general- more expensive than Banyan MINS and more complex to control.

Due to the advent of MINs, much research has been devoted to the investigation of their performance under various configurations and traffic conditions, and proposals have been made for the improvement of their performance. The main aspects that have been considered in these works are the buffer size of switching elements (e.g. [5], [6], [7]), MIN size (number of stages – e.g. [7], [8]), traffic patterns (including uniform vs. hotspot e.g. [9], [10], [7], [11], [12] and unicast vs. broadcast/multicast e.g. [18], [13]), and packet priorities (e.g. [10], [14], [11]). Performance evaluation has followed two distinct paths, the first one employing analytical methods such as Markov chains, queuing theory and Petri nets, while the second path uses simulation. Architectural issues (e.g. multilayer configurations [15] and wiring [16]) and routing algorithms (e.g. [17]) have also been considered in research efforts.

MIN performance under hotspot traffic and multiple priorities is receiving increasing attention, due to their correspondence with traffic patterns in real-world systems. Packet priority is a common issue in networks, arising when some packets need to be offered better quality of service than others. Packets with real-time requirements (e.g. from streaming media) vs. non real-time packets (e.g. file transfer), and out-of-band data vs. ordinary TCP traffic [19] are two examples of such differentiations. On the other hand, hotspot traffic is a typical situation when a server is deployed in some environment and clients access it frequently to obtain data and services, or when multiple network devices are interconnected via trunk ports. Insofar, however, the joint effect of packet priorities and hotspot traffic on the performance of MINs has not received adequate research attention. [10] and [11] are two works that have reported on this issue, but discuss an extreme hotspot situation, where all inputs send traffic to a specific output link and, additionally, all high-priority traffic is sent by a single input. Moreover, the MINs considered in these works are single-buffered, while more recent works (e.g. [20] and [21]) have shown that using double buffering or asymmetric buffering leads to elevated performance.

In this paper we examine performance aspects of dual-priority MINs under hotspot traffic conditions, considering different rates of offered load. We additionally take into account the differences in the performance of the MIN outputs under hotspot traffic identified in [22], according to which the performance of each output depends on the amount of overlapping that the path to the specific output has with the path to the hotspot output. We present metrics for the two most important network performance factors, namely throughput, delay and we also calculate and present the performance in terms of the Universal performance factor introduced in [6], which combines throughput and delay into...
a single metric, allowing the designer to express the perceived importance of each individual factor through weights.

The rest of this paper is organized as follows: in section 2 we briefly analyze the operation a Delta Network operating under hotspot traffic conditions and natively supporting 2-class routing traffic. Subsequently, in section 3 we introduce the published performance criteria and parameters related to this network. Section 4 presents the results of our performance analysis, which has been conducted through simulation experiments, while section 5 concludes the paper and outlines future work.

II. ANALYSIS OF 2-CLASS PRIORITY DELTA NETWORKS UNDER HOTSPOT ENVIRONMENT

A Multistage Interconnection Network (MIN) is generally defined as a network interconnecting a group of \( N \) inputs to a group of \( M \) outputs using several stages of small size Switching Elements (SEs). Each SE has a number of input and output links (this number is called the degree of the SE) and is followed (or preceded) by link states. MINs with the Banyan property are defined in [1] and are characterized by the fact that there is exactly a unique path from each source (input) to each sink (output). A Banyan MIN of size \( N \times N \) (i.e. connecting \( N \) inputs to \( N \) outputs) can be constructed by \( n = \log_2 N \) stages of \( (c \times c) \) SEs, where \( c \) is the degree of the SEs. At each stage there are exactly \( N/c \) SEs. An example MIN of size 8x8 is illustrated in Fig. 1. This MIN is assumed to operate under the following conditions at each input and output links (this number is called the degree of the SE) and is followed (or preceded) by link states. MINs with the Banyan property are defined in [1] and are characterized by the fact that there is exactly a unique path from each source (input) to each sink (output). A Banyan MIN of size \( N \times N \) (i.e. connecting \( N \) inputs to \( N \) outputs) can be constructed by \( n = \log_2 N \) stages of \( (c \times c) \) SEs, where \( c \) is the degree of the SEs. At each stage there are exactly \( N/c \) SEs. An example MIN of size 8x8 is illustrated in Fig. 1. This MIN is assumed to operate under the following conditions at each stage: 

- **output 0**, which is the output hotspot.
- **output 1**, which is the output adjacent to the hotspot output. Packets directed to this output have to contend with packets addressed to the hotspot output at all stages of the MIN, and they are free of such contention only when traversing the output link.

![Figure 1. An 8x8 delta-2 network with hotspot traffic](image-url)

- **outputs 2 and 3**, which are free of contention with packets addressed to the hotspot output when they traverse the last stage of the MIN. These outputs are termed as **Cold-1**, since they are free of contention with hotspot traffic for one stage.
- **outputs 4-7**, which are free of contention with packets addressed to the hotspot output when they traverse the last two stages of the network and thus are termed as **Cold-2**.

Generalizing, in an \( i \)-stage MIN, its output ports can be classified into the following \((i+1)\) zones: hotspot, adjacent, and cold-\( j \) \((1 \leq i \leq i-1)\).

Regarding priority support for the MIN depicted in Fig. 1, we can observe that individual queues have been added for both high and low priority packets. Thus, each SE has two transmission queues per link, with one queue dedicated to high priority packets and the other dedicated to low priority ones. Each queue is assumed to have two buffer positions for incoming packets.

Summarizing the above, a dual-priority, finite-buffered MIN is assumed to operate under the following conditions at hotspot environment:

- Routing is performed in a pipeline manner, meaning that the routing process occurs in every stage in parallel. Internal clocking results in synchronously operating switches in a slotted time model [23], and all SEs have deterministic service time.

- At each input of the network only one packet can be accepted within a time slot. All packets in input ports contain both the data to be transferred and the routing tag. The priority of each packet is indicated through a priority bit in the packet header. Under the dual-priority mechanism, when applications enter a packet to the network they specify its priority, designating it either as high or low. The offered load in all inputs of the network is uniform, all packets have the same size and the arrivals are independent of each other.

- There is a FIFO buffer in front of each SE enabling the packets of a message to be stored until they can be forwarded to the succeeding stage in the network.

- The backpressure mechanism deals with packets directed toward full buffers of the next stage, forcing them to stay in their current stage until the destination/s become available, so that no packets are lost inside the MIN.

- Under the dual-priority scheme, the SE considers all its links, examining for each one of them firstly the high priority queue. If this is not empty, it transmits the first packet towards the successive MIN stage; the low priority queue is checked only if the corresponding high priority queue is empty. In all cases, at most one packet per link (upper or lower) of a SE will be forwarded for each pair of high and low priority queues to the next stage. Conflicts between packets are solved randomly with equal probabilities.

- There is an initial fraction \( f_h \) of the total offered load \( \lambda \) that routed to the single hotspot output port. This fraction is exclusively low-priority traffic. The remaining packets, i.e. \( \lambda (1-f_h) \) are both high- and low-priority packets and are uniformly distributed across all destinations. That means
every output of the network except for hotspot has an equal probability of being one of the destinations of a packet. Note also that this input rate \( \lambda^*(1- f_{bs}) \) is addressed to all outputs, including the hotspot one, thus an additional load of \([\lambda^*(1- f_{bs})]/N\) is routed towards the hotspot output (including high- and low-priority packets).

- Packets are removed from their destinations immediately upon arrival, thus packets cannot be blocked at the last stage.

III. PERFORMANCE EVALUATION PARAMETERS AND METHODOLOGY

The following major parameters affect the performance of the test-bed dual priority Delta Network under a single hotspot environment.

- **Buffer size** \( b \) of a high or low priority queue is the maximum number of such packets that the corresponding input buffer of a SE can hold. In this paper we consider a double-buffered Delta Network, where \( b=2 \). We note here that the particular buffer size has been chosen since it has been reported [6] to provide optimal overall network performance: indeed, [6] documents that for smaller buffer sizes \( 1 \) the network throughput drops due to high blocking probabilities, whereas for higher buffer sizes \( 4 \) and \( 8 \) packet delay increases significantly (and the SE hardware cost also raises).

- **Offered load** \( \lambda \) is the steady-state fixed probability of arriving packets at each queue on inputs. In our simulation \( \lambda \) is assumed to be 0.1, 0.2, ..., 0.9, 1. \( \lambda \) can be further broken down to \( \lambda_{hs}, \lambda_{hp}, \lambda_{lp} \) which represent the arrival probability of the initial hotspot traffic, and the high and low priority traffic of the rest offered load respectively. It holds that \( \lambda_{hs} + \lambda_{hp} + \lambda_{lp} = 1 \).

- **Network size** \( n \), where \( n=\log_2 N \), is the number of stages of an \( N \times N \) Delta Network. In our simulation \( n \) is assumed to be \( n=6 \).

- **Hotspot fraction** \( f_{hs} \) is the fraction of the initial hotspot traffic which is considered to be \( f_{hs} = 0.05 \). We fix \( f_{hs} \) to this value, since using a higher value for a network of this size would lead to quick saturation of the paths to the hotspot output.

- **Ratio of high priority packets** \( r_{hp} \), is the ratio of high priority offered load for the normal traffic – i.e. excluding the traffic addressed to the initial hotspot - which is uniformly distributed among all output ports and it is assumed to be \( r_{hp} = 0.20 \). This ratio is generally adopted in works considering multiple priorities ([10], [14], [11]). Consequently,

\[
\lambda_{hs} = f_{hs} \lambda, \\
\lambda_{hp} = r_{hp} (1-f_{hs}) \lambda, \\
\lambda_{lp} = (1-r_{hp}) (1-f_{hs}) \lambda.
\]

Aiming to analyze the performance evaluation of a \( N \times N \) Delta Network with \( n=\log_2 N \) intermediate stages of \( (exc) \) SEs, the following metrics are used. Let \( T \) be a relatively large time period divided into \( u \) discrete time intervals \( (\tau_1, \tau_2, ..., \tau_u) \).

**Average throughput** \( Th_{av}(zone) \) of a specific output zone of MIN, where \( zone=\{\text{hotspot, adjacent, cold-1, ..., cold-(n-1)}\} \) is the average number of packets accepted by all destination ports of this zone per network cycle. Formally, \( Th_{av}(zone) \) is defined as

\[
Th_{av}(zone) = \lim_{u \rightarrow \infty} \sum_{i=1}^{u} n_{zone}(i)/u
\]

where \( n_{zone}(i) \) denotes the total number of packets routed to this specific output zone that reach their destinations during the \( i^{th} \) time interval.

**Normalized throughput** \( Th(zone) \) of a specific output zone of MIN is the ratio of the corresponding average throughput \( Th_{av}(zone) \) to the total number of output ports \( N(zone) \). Formally, \( Th(zone) \) can be expressed by

\[
Th(zone) = \frac{Th_{av}(zone)}{N(zone)}
\]

Relative normalized throughput of hotspot traffic \( RTh_{hs} \) is the normalized throughput \( Th(hotspot) \) of the hotspot output port divided by the corresponding ratio of packets on all input ports which are routed to single hotspot output port.

\[
RTh_{hs} = \frac{Th(hotspot)}{N \ast f_{hs} \ast (1 - f_{lp} \ast (1 - f_{hs}))}
\]

Relative normalized throughput of high priority traffic \( RTh_{hp} \) is the normalized throughput \( Th_{hp} \) of high priority packets routed to all output zones divided by the corresponding ratio of high priority packets on input ports.

\[
RTh_{hp} = \frac{Th_{hp}}{r_{hp} \ast (1 - f_{hs})}
\]

We do not report different \( RTh_{hp} \) for each zone, since our experiments have shown that this parameter is not affected by the zone when the MIN operates under the parameter ranges listed above.

**Relative normalized throughput** of low priority traffic \( RTh_{lp}(zone) \) routed to a specific zone of output ports is the normalized throughput \( Th_{lp}(zone) \) of such packets divided by the corresponding ratio of low priority packets on input ports.

\[
RTh_{lp}(zone) = \frac{Th_{lp}(zone)}{1 - r_{hp} \ast (1 - f_{hs})}
\]

**Average packet delay** \( D_{av}(zone) \) of packets routed to specific output zone of MIN is the average time the these packets spend to pass through the network. Formally, \( D_{av}(zone) \) is expressed by
where \( n(z, u) \) denotes the total number of packets accepted within \( u \) time intervals, while \( t_i(z, i) \) represents the delay of the \( i \)th packet to traverse from an input port to a port of the specific output zone. We consider \( t_i(z, i) = t_w(z, i) + t_{tr}(z, i) \) where \( t_w(z, i) \) denotes the total queuing delay for \( i \)th packet waiting at each stage for the availability of an empty buffer at the next stage queue of the network. The second term \( t_{tr}(z, i) \) denotes the transmission delay for \( i \)th packet at each stage of the network, that is just \( n \ast nc \), where \( n \) is the number of stages and \( nc \) is the network cycle.

**Normalized packet delay** \( D(z) \) is the ratio of the \( D_{avg}(z) \) to the minimum packet delay which is simply the transmission delay \( n \ast nc \) (i.e. zero queuing delay). Formally, \( D(z) \) can be defined as

\[
D(z) = \frac{D_{avg}(z)}{n \ast nc}
\]  

(7)

**Universal performance** \( U(z) \) is defined through a formula involving the two major above normalized factors, namely \( D(z) \) and \( RTh(z) \): the performance of a zone \( z \) of Delta Network is considered optimal when \( D(z) \) is minimized and \( RTh(z) \) is maximized, thus the formula for computing the universal factor arranges so that the overall performance metric follows that rule. Formally, \( U(z) \) can be expressed by

\[
U(z) = \sqrt[2]{\frac{D(z)}{2} + \frac{1}{RTh(z)^2}}
\]  

(8)

It is obvious that, when the packet delay factor becomes smaller at/and throughput factor becomes larger the universal performance factor \( U \) becomes smaller. Consequently, as the universal performance factor \( U \) becomes smaller, the performance of Delta Network is considered to improve. Because the above factors (parameters) have different measurement units and scaling, we normalize them to obtain a reference value domain.

**4. Simulation and Performance Results**

The overall network performance of finite buffered MINs under hotspot environment was evaluated by developing a special-purpose simulator in C++, capable to handle dual priority traffic. This type of modeling [24] using simulation experiments was applied due to the complexity of the mathematical model [25], stemming from the combination of multi-priority with hotspot traffic. Several input parameters such as the buffer-length, the number of input and output ports, the initial hotspot fraction, and the ratio of high priority packets were considered. Internally, each SE was modelled by four non-shared buffer queues, the first two dedicated for high priority packets, and the other two for low priority ones, where buffer operation was based on the FCFS principle. All simulation experiments were performed at packet level, assuming fixed-length packets transmitted in equal-length time slots, where the slot was the time required to forward a packet from one stage to the successive. The contention between two packets were resolved by favoring the packet transmitted from the queue in which high priority packets were stored in; contentions between equal-priority packets were resolved by choosing randomly one of the packets for transmission, whereas the other packet was blocked.

Metrics such as packet throughput, and packet delay were collected. We performed extensive simulations to validate our results. All statistics obtained from simulation running for \( 10^5 \) clock cycles. The number of simulation runs was adjusted to ensure a steady-state operating condition for the MIN. There was a stabilization phase to allow the network to reach a steady state, by discarding the data from the first \( 10^3 \) network cycles, before initiating metrics collection.

**A. Simulator validation**

Since no other simulator/model supporting dual priority traffic under hotspot environment has been reported insofar in the literature, we validated our simulator against those that have been made available; i.e. single-priority under hotspot environment and dual-priority under uniform traffic conditions.

In the case of hotspot environment, the measurements reported in table 1 of [26] and those obtained by our simulator in the marginal case of single-priority traffic, where \( r_{up} = 0.10, f_{hs} = 0.10, \) and \( N=8 \), have found to be in close agreement (all differences were less than 2%).

On the other hand, the priority mechanism was tested under uniform traffic conditions; this was done by setting the parameter \( f_{hs} = 0 \). We compared our measurements against those obtained from Shabtai's Model reported in [10], and
have found that both results are in close agreement (the maximum difference was only 3.8%).

Fig. 2 illustrates this comparison, involving the total normalized throughput for all packets (both high and low priority) of a dual-priority, single-buffered, 6-stage MIN vs. the ratio of high priority packets under full offered load.

![Figure 2](image1)

**Figure 2.** Total normalized throughput of a dual-priority, single-buffered, 6-stage MIN

**B. Dual-priority MINs under Hotspot Environment**

In this paper we extend the study of hotspot environment in MINs by considering dual-priority SEs in order to support varying quality of service for packets and by using double-buffered queues to improve overall network performance.

![Figure 3](image2)

**Figure 3.** Normalized throughput of a single priority, double-buffered, 6-stage MIN under hotspot traffic

Fig. 3 depicts the relative normalized throughput of a single-priority, double-buffered, 6-stage Delta Network for single hotspot output port, as well as the cold-3 and cold-5 zones in comparison with the normalized throughput of the corresponding MIN configuration under uniform traffic conditions, when the initial hotspot traffic is set to $f_{ha}=0.05$. It is obvious that the non-uniform traffic causes a serious traffic congestion problem not only to the single hotspot output port but also to the zones which are more close to it. According to fig. 3, the performance degradation of both hotspot and cold-3 zone is approximately 58.5%, while the cold-5 zone exhibits improved performance, mainly owing to the fact that it has a lighter load (recall that a ratio equal to $f_{ha}$ is addressed to the hotspot output, and this is subtracted from the load of other outputs).

![Figure 4](image3)

**Figure 4.** Normalized throughput of a dual-priority, double-buffered, 6-stage MIN under hotspot traffic

As a response to the tree saturation problem, a dual-priority MIN configuration can offer better quality-of-service to some applications by prioritizing their packets. According to Fig. 4, the relative normalized throughput of high priority packets approaches the optimal value $R_{Th_{hp}}=1$, when the initial hotspot traffic is $f_{ha}=0.05$ and the ratio of high priority packets is $r_{hp}=0.20$. Recall from previous section the relative normalized throughput of high priority packets is evaluated by collecting measurements on all output ports, showing that the gain is higher for the single hotspot output port and the zones which are more close to it (since these zones exhibit the most acute performance deterioration under hotspot traffic). We can also notice that the throughput of low priority traffic for hotspot and cold-3 zones is slightly improved against the respective performance in Fig. 3: this can be attributed to the introduction of the additional buffers in the SEs (recall that SEs have distinct buffers for high- and low-priority packets). The cold-5 zone, on the other hand, exhibits a slight deterioration towards the full input load when compared to Fig. 3, with the performance curve converging to the Single Priority/Uniform curve. This is owing to the fact that at this load range, the network has many high-priority packets to serve, thus the service offered to low priority packets is degraded.

Figs. 5 and 6 represent the findings for the normalized packet delay of single- and dual-priority MINs, under hotspot environment. Again we can observe that high-priority packets obtain service close to the optimal one, at all offered load setups. It is worth noting that normalized packet delay of hotspot traffic is effectively double than the delay of the cold-3 zone, while the divergence between the two zones regarding relative normalized throughput is negligible at both configurations. Finally, the delay for packets routed to cold-5 zone is scientifically smaller than the delay of packets.
routed to cold-3 zone. The small drop in the low priority packet delay towards the full load area in Fig. 6 is owing to the fact that --for that area-- a number of low priority packets is not accepted for entrance in the network, due to buffer unavailability at the first MIN stage.

Similarly, figs. 7 and 8 depict the behavior of the universal performance factor of single- and dual-priority MINs, under hotspot traffic conditions. We can observe that high-priority packets obtain again service close to the optimal zero, under full offered load. We can also notice that the difference in the delay factor between zones hotspot and cold-3 is reflected in the Universal Performance Factor (although both zones have the same throughput), and that zone cold-5 exhibits considerably better performance for loads $\lambda > 0.2$.

V. UNIVERSAL PERFORMANCE OF A DUAL-PRIORITY, DOUBLE-BUFFERED, 6-STAGE MIN UNDER HOTSPOT TRAFFIC

VI. CONCLUSIONS

In this paper we have examined the performance of MINs natively supporting two priorities, when these operate under hotspot traffic conditions. Our findings that when the hotspot conditions are not extreme and the high priority packet ratio is moderate (20%), high priority packets receive almost optimal quality of service, whereas the QoS offered to low priority packets varies, depending on the zone they are addressed to. It is also interesting that while throughput for some zones is found to be identical, the same zones exhibit variations of behavior regarding the delay metric. In all cases, performance indicators of low-priority packets for
zones that are “close” to the hotspot output appear to quickly deteriorate even for light loads \((\lambda \geq 0.3)\), whereas low-priority packets addressed to zones “far” from the hotspot output exhibit a performance similar to that of MINs under uniform input load.

Future work will include further experimentation with operating parameters of the MIN, including the overall network size, the high/low priority packet ratio and the hotspot/normal traffic ratio. The introduction of an adaptive scheme, altering buffer allocation to different priority classes according to current traffic load and high/low priority ratios will be investigated as well.

VII. REFERENCES